Universität Stuttgart

Institut für Maschinelle Sprachverarbeitung

Sabine Schulte im Walde

Feature-based Compositionality Ratings for Noun Compounds

Motivation, Study & Research Questions

Lexical Resources for Semantic Evaluation

- **Starting point**:
  Developing computational models to predict degrees of compositionality for multi-word expressions typically goes hand in hand with creating or using reliable lexical resources as gold standards for formative intrinsic evaluation.
- **Problems**:
  - How much vary both the gold standards and the prediction models according to properties of the targets within the lexical resources?
  - Potential skewness hinders us from a generalised assessment of models.
- **Focus**: English and German noun compounds
- **Contributions**:
  - Novel collection of compositionality ratings for 1,099 German noun compounds, where we asked the human judges to provide compound and constituent properties before judging the compositionality
  - Series of analyses on rating distributions and interactions with compound and constituent properties

Multiword Expressions & Noun Compounds

- **Multiword expressions**:
  combinations of words with some degree of idiosynchracy, i.e., the meaning of the combination is not entirely (or even not at all) predictable from the meanings of the constituents [Sag et al., 2002, Baldwin and Kim, 2010]
- **Noun compounds**: compositions of modifier and nominal head constituents
- **Compositionality**: meaning contributions of constituents to compound meaning; strength of semantic relatedness: compounds ↔ constituents
- **Computational task & models**:
  - Task: predict the degree of compound compositionality as a whole/phrase and with regard to its constituents
  - Models: textual/multi-modal vector-space models (VSMs)

META-LEVEL RESEARCH QUESTIONS

- **To what extent should we aim for an even distribution of human ratings across a pre-specified scale?**
- **To what extent should we take into account properties of targets when creating a novel resource and when using a resource?**

Datasets & Computational Models

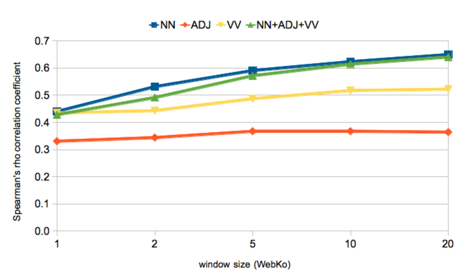
Datasets of Noun Compound Compositionality

- **REDDY-N** (English) [Reddy et al., 2011]
  - WordNet-based heuristic: a compound is considered compositional with regard to a constituent if the constituent represents a hypernym of the compound or is used in the definition, e.g., *swimming pool*
  - 90 noun-noun compounds; scale [0,5]
- **CONCRETE-NN** (German) [von der Heide and Borgwaldt, 2009, Schulte im Walde et al., 2013]
  - 244 depictable noun-noun compounds; scale [1,7]
- **G$_h$OST-NN** (German) [Schulte im Walde et al., 2016]
  - **G$_h$OST-NN/S**: 20 × 9 = 180 compounds randomly extracted from corpus but balanced for modifier productivity (low/mid/high) and head ambiguity (1/2/>2)
  - **G$_h$OST-NN/XL**: 868 compounds, after adding all compounds with the same modifiers and heads as in G$_h$OST/S

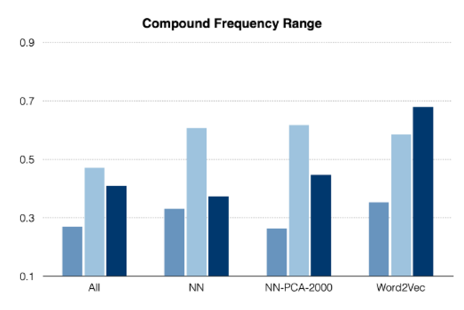
compound examples	**mean ratings**		
	compound	modifier	head
climate change	**4.97**±0.18	**4.90**±0.30	**4.83**±0.38
couch potato	**1.41**±1.03	**3.27**±1.48	**0.34**±0.66
crocodile tears	**1.25**±1.09	**0.19**±0.47	**3.79**±1.05
melting pot	**0.54**±0.63	**1.00**±1.15	**0.48**±0.63
night owl	**1.93**±1.27	**4.47**±0.88	**0.50**±0.82
Ahornblatt (maple leaf)	**6.03**±1.49	**5.64**±1.63	**5.71**±1.70
Fliegenpilz (toadstool, lit. fly mushroom)	**2.00**±1.20	**1.93**±1.28	**6.55**±0.63
Flohmarkt (flea market)	**2.31**±1.65	**1.50**±1.22	**6.03**±1.50
Löwenzahn (dandelion, lit. lion tooth)	**1.66**±1.54	**2.10**±1.84	**2.23**±1.92
Windlicht (storm lamp, lit. wind light)	**3.52**±2.08	**3.07**±2.12	**4.27**±2.36

Vector-Space Models Predicting Compositionality

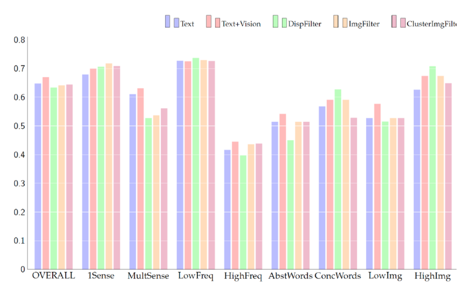
- **Basis**: vectors-space representations for compounds and constituents
- **Relatedness**: mathematical distance measure between vectors of compounds and vectors of their modifier and head constituents
- **Compositionality**: VSM relatedness ∼ compositionality
- **Evaluation**: Spearman's rank-order correlation coefficient ρ relating predicted distances ∼ compositionality scores



CONCRETE-NN:
window-based models



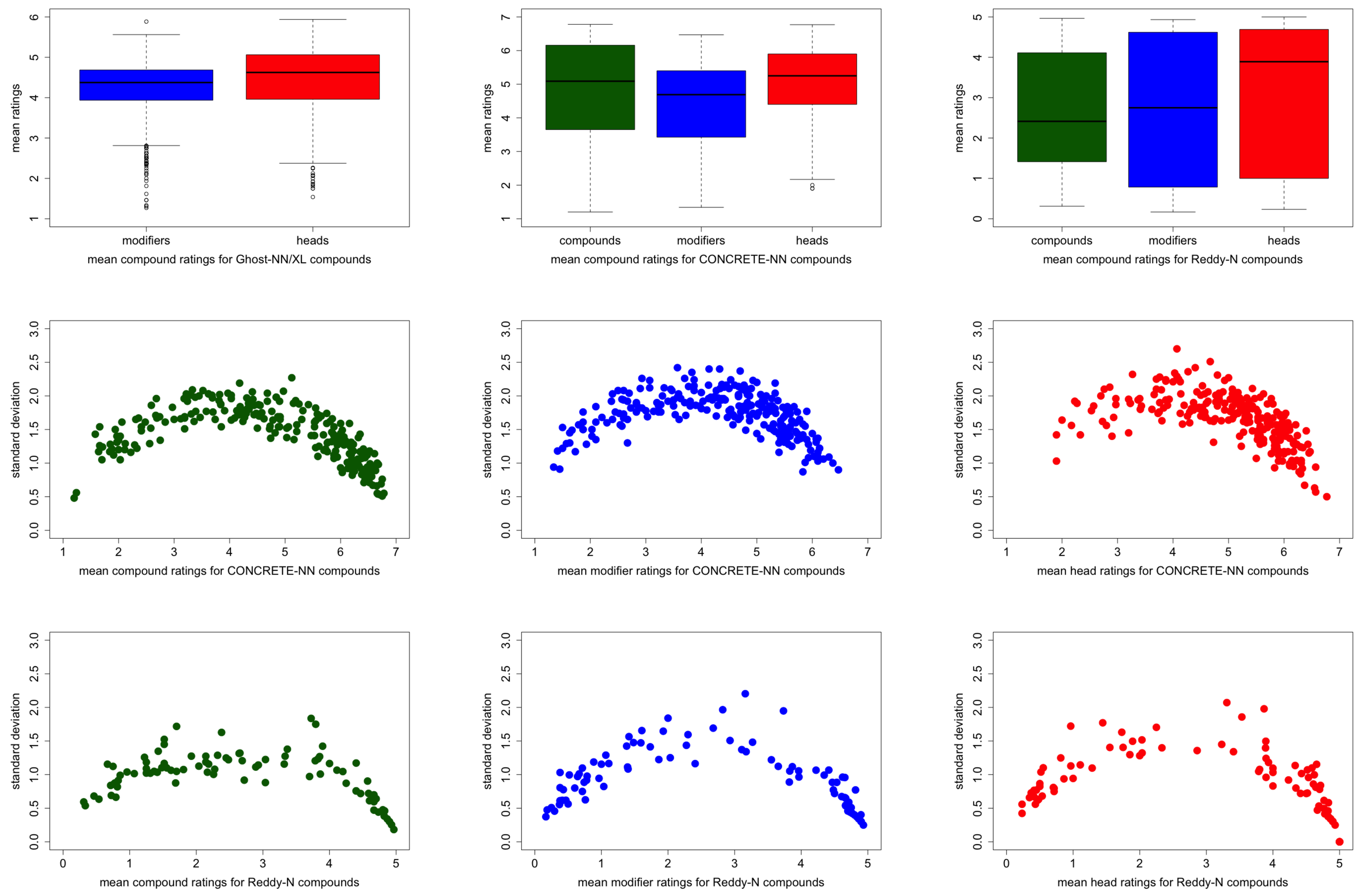
REDDY-N:
count vs. predict/reduced models



GHOST-NN:
multimodal models (text+images)

Analyses

Compositionality Rating Distributions



→ Datasets are skewed towards certain ranges of compositionality in different ways.

Compositionality and Target Properties

		freq	prod	amb	hyp	conc
CONCRETE-NN	compound	-.075	–	–	.424	.113
	modifier	.080	.164	-.157	–	.079
	head	-.147	-.178	-.279	.689	.228
GHOST-NN/XL	modifier	-.088	-.023	-.231	–	.119
	head	-.202	-.204	-.356	.692	.171
REDDY-N	compound	.579	–	–	–	.615
	modifier	.547	.471	.172	–	.318
	head	.454	.484	.224	–	.622

→ Some datasets exhibit strong correlations between compound and constituent ratings, and moderate correlations between compositionality ratings and corpus-based frequencies and productivity scores.

META-LEVEL SUGGESTIONS

- Balance your targets across frequency ranges as the minimally required target property, because we know that target frequency has generally a strong influence on language processing and comprehension.
- Assess models not only on the full dataset, but also with regard to subsets of targets with coherent task-relevant properties.